SemEval-2010 Task 13: TempEval-2

Marc Verhagen[†], **Roser Saurí** [‡], **Tommaso Caselli**[*] **and James Pustejovsky**[†]

† Computer Science Department, Brandeis University, Massachusetts, USA
‡Barcelona Media, Barcelona, Spain ∗ ILC-CNR, Pisa, Italy

marc@cs.brandeis.edu roser.sauri@barcelonamedia.org
tommaso.caselli@ilc.cnr.it jamesp@cs.brandeis.edu

Abstract

Tempeval-2 comprises evaluation tasks for time expressions, events and temporal relations, the latter of which was split up in four sub tasks, motivated by the notion that smaller subtasks would make both data preparation and temporal relation extraction easier. Manually annotated data were provided for six languages: Chinese, English, French, Italian, Korean and Spanish.

1 Introduction

The ultimate aim of temporal processing is the automatic identification of all temporal referring expressions, events and temporal relations within a text. However, addressing this aim is beyond the scope of an evaluation challenge and a more modest approach is appropriate.

The 2007 SemEval task, TempEval-1 (Verhagen et al., 2007; Verhagen et al., 2009), was an initial evaluation exercise based on three limited temporal ordering and anchoring tasks that were considered realistic both from the perspective of assembling resources for development and testing and from the perspective of developing systems capable of addressing the tasks.[1]

TempEval-2 is based on TempEval-1, but is more elaborate in two respects: (i) it is a multilingual task, and (ii) it consists of six subtasks rather than three.

In the rest of this paper, we first introduce the data that we are dealing with. Which gets us in a position to present the list of task introduced by TempEval-2, including some motivation as to why we feel that it is a good idea to split up temporal relation classification into sub tasks. We proceed by shortly describing the data resources and their creation, followed by the performance of the systems that participated in the tasks.

2 TempEval Annotation

The TempEval annotation language is a simplified version of TimeML.[2] using three TimeML tags: TIMEX3, EVENT and TLINK.

TIMEX3 tags the time expressions in the text and is identical to the TIMEX3 tag in TimeML. Times can be expressed syntactically by adverbial or prepositional phrases, as shown in the following example.

(1) a. on Thursday
b. November 15, 2004
c. Thursday evening
d. in the late 80's
e. later this afternoon

The two main attributes of the TIMEX3 tag are TYPE and VAL, both shown in the example (2).

(2) *November 22, 2004*
`type="DATE" val="2004-11-22"`

For TempEval-2, we distinguish four temporal types: TIME (*at 2:45 p.m.*), DATE (*January 27, 1920, yesterday*), DURATION (*two weeks*) and SET (*every Monday morning*). The VAL attribute assumes values according to an extension of the ISO 8601 standard, as enhanced by TIMEX2.

Each document has one special TIMEX3 tag, the Document Creation Time (DCT), which is interpreted as an interval that spans a whole day.

The EVENT tag is used to annotate those elements in a text that describe what is conventionally referred to as an *eventuality*. Syntactically, events are typically expressed as inflected verbs, although event nominals, such as "crash" in *killed by the crash*, should also be annotated as EVENTs. The most salient event attributes encode tense, aspect, modality and polarity information. Examples of some of these features are shown below:

[1] The Semeval-2007 task was actually known simply as TempEval, but here we use Tempeval-1 to avoid confusion.

[2] See `http://www.timeml.org` for language specifications and annotation guidelines

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 57–62,
Uppsala, Sweden, 15-16 July 2010.

(3) should have *bought*
`tense="PAST" aspect="PERFECTIVE"`
`modality="SHOULD" polarity="POS"`

(4) did not *teach*
`tense="PAST" aspect="NONE"`
`modality="NONE" polarity="NEG"`

The relation types for the TimeML TLINK tag form a fine-grained set based on James Allen's interval logic (Allen, 1983). For TempEval, the set of labels was simplified to aid data preparation and to reduce the complexity of the task. We use only six relation types including the three core relations BEFORE, AFTER, and OVERLAP, the two less specific relations BEFORE-OR-OVERLAP and OVERLAP-OR-AFTER for ambiguous cases, and finally the relation VAGUE for those cases where no particular relation can be established.

Temporal relations come in two broad flavours: anchorings of events to time expressions and orderings of events. Events can be anchored to an adjacent time expression as in examples 5 and 6 or to the document creation time as in 7.

(5) Mary *taught*$_{e1}$ on *Tuesday morning*$_{t1}$
OVERLAP(e1,t1)

(6) They cancelled the *evening*$_{t2}$ *class*$_{e2}$
OVERLAP(e2,t2)

(7) Most troops will *leave*$_{e1}$ Iraq by August of 2010. AFTER(e1,dct)
The country *defaulted*$_{e2}$ on debts for that entire year. BEFORE(e2,dct)

In addition, events can be ordered relative to other events, as in the examples below.

(8) The President *spoke*$_{e1}$ to the nation on Tuesday on the financial crisis. He had *conferred*$_{e2}$ with his cabinet regarding policy the day before. AFTER(e1,e2)

(9) The students *heard*$_{e1}$ a *fire alarm*$_{e2}$.
OVERLAP(e1,e2)

(10) He *said*$_{e1}$ they had *postponed*$_{e2}$ the meeting.
AFTER(e1,e2)

3 TempEval-2 Tasks

We can now define the six TempEval tasks:

A. Determine the extent of the time expressions in a text as defined by the TimeML TIMEX3 tag. In addition, determine value of the features TYPE and VAL.

B. Determine the extent of the events in a text as defined by the TimeML EVENT tag. In addition, determine the value of the features CLASS, TENSE, ASPECT, POLARITY, and MODALITY.

C. Determine the temporal relation between an event and a time expression in the same sentence. This task is further restricted by requiring that either the event syntactically dominates the time expression or the event and time expression occur in the same noun phrase.

D. Determine the temporal relation between an event and the document creation time.

E. Determine the temporal relation between two main events in consecutive sentences.

F. Determine the temporal relation between two events where one event syntactically dominates the other event.

Of these tasks, C, D and E were also defined for TempEval-1. However, the syntactic locality restriction in task C was not present in TempEval-1.

Task participants could choose to either do all tasks, focus on the time expression task, focus on the event task, or focus on the four temporal relation tasks. In addition, participants could choose one or more of the six languages for which we provided data: Chinese, English, French, Italian, Korean, and Spanish.

We feel that well-defined tasks allow us to structure the workflow, allowing us to create task-specific guidelines and using task-specific annotation tools to speed up annotation. More importantly, each task can be evaluated in a fairly straightforward way, contrary to for example the problems that pop up when evaluating two complex temporal graphs for the same document. In addition, tasks can be ranked, allowing systems to feed the results of one (more precise) task as a feature into another task.

Splitting the task into substask reduces the error rate in the manual annotation, and that merging the different sub-task into a unique layer as a post-processing operation (see figure 1) provides better

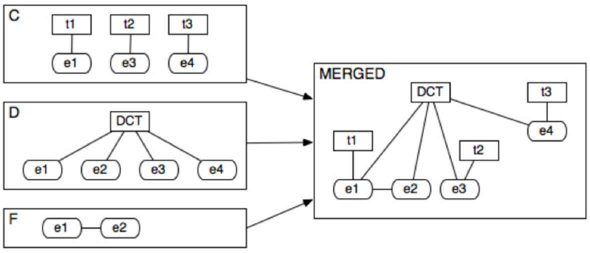

Figure 1: Merging Relations

and more reliable results (annotated data) than doing a complex task all at once.

4 Data Preparation

The data for the five languages were prepared independently of each other and do not comprise a parallel corpus. However, annotation specifications and guidelines for the five languages were developed in conjunction with one other, in many cases based on version 1.2.1 of the TimeML annotation guidelines for English[3]. Not all corpora contained data for all six tasks. Table 1 gives the size of the training set and the relation tasks that were included.

language	tokens	C	D	E	F	X
Chinese	23,000	✓	✓	✓	✓	
English	63,000	✓	✓	✓	✓	
Italian	27,000	✓	✓	✓		
French	19,000					✓
Korean	14,000					
Spanish	68,000	✓	✓			

Table 1: Corpus size and relation tasks

All corpora include event and timex annotation. The French corpus contained a subcorpus with temporal relations but these relations were not split into the four tasks C through F.

Annotation proceeded in two phases: a dual annotation phase where two annotators annotate each document and an adjudication phase where a judge resolves disagreements between the annotators. Most languages used BAT, the Brandeis Annotation Tool (Verhagen, 2010), a generic web-based annotation tool that is centered around the notion of annotation tasks. With the task decomposition allowed by BAT, it is possible to structure the complex task of temporal annotation by splitting it up in as many sub tasks as seems useful. As such, BAT was well-suited for TempEval-2 annotation.

We now give a few more details on the English and Spanish data, skipping the other languages for reasons that will become obvious at the beginning of section 6.

The English data sets were based on TimeBank (Pustejovsky et al., 2003; Boguraev et al., 2007), a hand-built gold standard of annotated texts using the TimeML markup scheme.[4] However, all event annotation was reviewed to make sure that the annotation complied with the latest guidelines and all temporal relations were added according to the Tempeval-2 relation tasks, using the specified relation types.

The data released for the TempEval-2 Spanish edition is a fragment of the Spanish TimeBank, currently under development. Its documents are originally from the Spanish part of the AnCora corpus (Taulé et al., 2008). Data preparation followed the annotation guidelines created to deal with the specificities of event and timex expressions in Spanish (Saurí et al., 2009a; Saurí et al., 2009b).

5 Evaluation Metrics

For the extents of events and time expressions (tasks A and B), precision, recall and the f1-measure are used as evaluation metrics, using the following formulas:

$$\begin{aligned}
precision &= tp/(tp + fp) \\
recall &= tp/(tp + fn) \\
f\text{-}measure &= 2 * (P * R)/(P + R)
\end{aligned}$$

Where *tp* is the number of tokens that are part of an extent in both key and response, *fp* is the number of tokens that are part of an extent in the response but not in the key, and *fn* is the number of tokens that are part of an extent in the key but not in the response.

For attributes of events and time expressions (the second part of tasks A and B) and for relation types (tasks C through F) we use an even simpler metric: the number of correct answers divided by the number of answers.

[3] See `http://www.timeml.org`.

[4] See `www.timeml.org` for details on TimeML, TimeBank is distributed free of charge by the Linguistic Data Consortium (`www.ldc.upenn.edu`), catalog number LDC2006T08.

6 System Results

Eight teams participated in TempEval-2, submitting a grand total of eighteen systems. Some of these systems only participated in one or two tasks while others participated in all tasks. The distribution over the six languages was very uneven: sixteen systems for English, two for Spanish and one for English and Spanish.

The results for task A, recognition and normalization of time expressions, are given in tables 2 and 3.

team	p	r	f	type	val
UC3M	0.90	0.87	0.88	0.91	0.83
TIPSem	0.95	0.87	0.91	0.91	0.78
TIPSem-B	0.97	0.81	0.88	0.99	0.75

Table 2: Task A results for Spanish

team	p	r	f	type	val
Edinburgh	0.85	0.82	0.84	0.84	0.63
HeidelTime1	0.90	0.82	0.86	0.96	0.85
HeidelTime2	0.82	0.91	0.86	0.92	0.77
JU_CSE	0.55	0.17	0.26	0.00	0.00
KUL	0.78	0.82	0.80	0.91	0.55
KUL Run 2	0.73	0.88	0.80	0.91	0.55
KUL Run 3	0.85	0.84	0.84	0.91	0.55
KUL Run 4	0.76	0.83	0.80	0.91	0.51
KUL Run 5	0.75	0.85	0.80	0.91	0.51
TERSEO	0.76	0.66	0.71	0.98	0.65
TIPSem	0.92	0.80	0.85	0.92	0.65
TIPSem-B	0.88	0.60	0.71	0.88	0.59
TRIOS	0.85	0.85	0.85	0.94	0.76
TRIPS	0.85	0.85	0.85	0.94	0.76
USFD2	0.84	0.79	0.82	0.90	0.17

Table 3: Task A results for English

The results for Spanish are more uniform and generally higher than the results for English. For Spanish, the f-measure for TIMEX3 extents ranges from 0.88 through 0.91 with an average of 0.89; for English the f-measure ranges from 0.26 through 0.86, for an average of 0.78. However, due to the small sample size it is hard to make any generalizations. In both languages, type detection clearly was a simpler task than determining the value.

The results for task B, event recognition, are given in tables 4 and 5. Both tables contain results for both Spanish and English, the first part of each table contains the results for Spanish and the next part the results for English.

team	p	r	f
TIPSem	0.90	0.86	0.88
TIPSem-B	0.92	0.85	0.88
team	p	r	f
Edinburgh	0.75	0.85	0.80
JU_CSE	0.48	0.56	0.52
TIPSem	0.81	0.86	0.83
TIPSem-B	0.83	0.81	0.82
TRIOS	0.80	0.74	0.77
TRIPS	0.55	0.88	0.68

Table 4: Event extent results

The column headers in table 5 are abbreviations for polarity (pol), mood (moo), modality (mod), tense (tns), aspect (asp) and class (cl). Note that the English team chose to include modality whereas the Spanish team used mood.

team	pol	moo	tns	asp	cl
TIPSem	0.92	0.80	0.96	0.89	0.66
TIPSem-B	0.92	0.79	0.96	0.89	0.66
team	pol	mod	tns	asp	cl
Edinburgh	0.99	0.99	0.92	0.98	0.76
JU_CSE	0.98	0.98	0.30	0.95	0.53
TIPSem	0.98	0.97	0.86	0.97	0.79
TIPSem-B	0.98	0.98	0.85	0.97	0.79
TRIOS	0.99	0.95	0.91	0.98	0.77
TRIPS	0.99	0.96	0.67	0.97	0.67

Table 5: Event attribute results

As with the time expressions results, the sample size for Spanish is small, but note again the higher f-measure for event extents in Spanish.

Table 6 shows the results for all relation tasks, with the Spanish systems in the first two rows and the English systems in the last six rows. Recall that for Spanish the training and test sets only contained data for tasks C and D.

Interestingly, the version of the TIPSem systems that were applied to the Spanish data did much better on task C compared to its English cousins, but much worse on task D, which is rather puzzling.

Such a difference in performance of the systems could be due to differences in annotation accurateness, or it could be due to some particularities of how the two languages express certain temporal

team	C	D	E	F
TIPSem	0.81	0.59	-	-
TIPSem-B	0.81	0.59	-	-
JU_CSE	0.63	0.80	0.56	0.56
NCSU-indi	0.63	0.68	0.48	0.66
NCSU-joint	0.62	0.21	0.51	0.25
TIPSem	0.55	0.82	0.55	0.59
TIPSem-B	0.54	0.81	0.55	0.60
TRIOS	0.65	0.79	0.56	0.60
TRIPS	0.63	0.76	0.58	0.59
USFD2	0.63	-	0.45	-

Table 6: Results for relation tasks

aspects, or perhaps the one corpus is more homogeneous than the other. Again, there are not enough data points, but the issue deserves further attention.

For each task, the test data provided the event pairs or event-timex pairs with the relation type set to NONE and participating systems would replace that value with one of the six allowed relation types. However, participating systems were allowed to not replace NONE and not be penalized for it. Those cases would not be counted when compiling the scores in table 6. Table 7 lists those systems that did not classify all relation and the percentage of relations for each task that those systems did not classify.

team	C	D	E	F
TRIOS	25%	19%	36%	31%
TRIPS	20%	10%	17%	10%

Table 7: Percentage not classified

A comparison with the Tempeval-1 results from Semeval-2007 may be of interest. Six systems participated in the TempEval-1 tasks, compared to seven or eight systems for TempEval-2. Table 8 lists the average scores and the standard deviations for all the tasks (on the English data) that Tempeval-1 and Tempeval-2 have in common.

		C	D	E
tempeval-1	average	0.59	0.76	0.51
	stddev	0.03	0.03	0.05
tempeval-2	average	0.61	0.70	0.53
	stddev	0.04	0.22	0.05

Table 8: Comparing Tempevals

The results are very similar except for task D, but if we take a away the one outlier (the NCSU-joint score of 0.21) then the average becomes 0.78 with a standard deviation of 0.05. However, we had expected that for TempEval-2 the systems would score better on task C since we added the restriction that the event and time expression had to be syntactically adjacent. It is not clear why the results on task C have not improved.

7 Conclusion

In this paper, we described the TempEval-2 task within the SemEval 2010 competition. This task involves identifying the temporal relations between events and temporal expressions in text. Using a subset of TimeML temporal relations, we show how temporal relations and anchorings can be annotated and identified in six different languages. The markup language adopted presents a descriptive framework with which to examine the temporal aspects of natural language information, demonstrating in particular, how tense and temporal information is encoded in specific sentences, and how temporal relations are encoded between events and temporal expressions. This work paves the way towards establishing a broad and open standard metadata markup language for natural language texts, examining events, temporal expressions, and their orderings.

One thing that would need to be addressed in a follow-up task is what the optimal number of tasks is. Tempeval-2 had six tasks, spread out over six languages. This brought about some logistical challenges that delayed data delivery and may have given rise to a situation where there was simply not enough time for many systems to properly prepare. And clearly, the shared task was not successful in attracting systems to four of the six languages.

8 Acknowledgements

Many people were involved in TempEval-2. We want to express our gratitude to the following key contributors: Nianwen Xue, Estela Saquete, Lotus Goldberg, Seohyun Im, André Bittar, Nicoletta Calzolari, Jessica Moszkowicz and Hyopil Shin.

Additional thanks to Joan Banach, Judith Domingo, Pau Giménez, Jimena del Solar, Teresa Suñol, Allyson Ettinger, Sharon Spivak, Nahed Abul-Hassan, Ari Abelman, John Polson, Alexandra Nunez, Virginia Partridge, , Amber Stubbs, Alex Plotnick, Yuping Zhou, Philippe Muller and

Irina Prodanof.

The work on the Spanish corpus was supported by a EU Marie Curie International Reintegration Grant (PIRG04-GA-2008-239414). Work on the English corpus was supported under the NSF-CRI grant 0551615, ”Towards a Comprehensive Linguistic Annotation of Language” and the NSF-INT-0753069 project ”Sustainable Interoperability for Language Technology (SILT)”, funded by the National Science Foundation.

Finally, thanks to all the participants, for sticking with a task that was not always as flawless and timely as it could have been in a perfect world.

References

James Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Bran Boguraev, James Pustejovsky, Rie Ando, and Marc Verhagen. 2007. Timebank evolution as a community resource for timeml parsing. *Language Resource and Evaluation*, 41(1):91–115.

James Pustejovsky, David Day, Lisa Ferro, Robert Gaizauskas, Patrick Hanks, Marcia Lazo, Roser Saurí, Andrew See, Andrea Setzer, and Beth Sundheim. 2003. The TimeBank Corpus. *Corpus Linguistics*, March.

Roser Saurí, Olga Batiukova, and James Pustejovsky. 2009a. Annotating events in spanish. timeml annotation guidelines. Technical Report Version TempEval-2010., Barcelona Media - Innovation Center.

Roser Saurí, Estela Saquete, and James Pustejovsky. 2009b. Annotating time expressions in spanish. timeml annotation guidelines. Technical Report Version TempEval-2010, Barcelona Media - Innovation Center.

Mariona Taulé, Toni Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the LREC 2008*, Marrakesh, Morocco.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proc. of the Fourth Int. Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*.

Marc Verhagen. 2010. The Brandeis Annotation Tool. In *Language Resources and Evaluation Conference, LREC 2010*, Malta.